Development of a Statewide Trauma Registry Using Multiple Linked Sources of Data

David E. Clark, M.D., Department of Surgery
Maine Medical Center, Portland, Maine

In order to develop a cost-effective method of injury surveillance and trauma system evaluation in a rural state, computer programs were written linking records from two major hospital trauma registries, a statewide trauma tracking study, hospital discharge abstracts, death certificates, and ambulance run reports. A general-purpose database management system, programming language, and operating system were used. Data from 1991 appeared to be successfully linked using only indirect identifying information. Familiarity with local geography and the idiosyncracies of each data source were helpful in programming for effective matching of records. For each individual case identified in this way, data from all available sources were then merged and imported into a standard database format. This inexpensive, population-based approach, maintaining flexibility for end-users with some database training, may be adaptable for other regions. There is a need for further improvement and simplification of the record-linkage process for this and similar purposes.

INTRODUCTION

The increasing availability of microcomputers has led to the development of trauma registries at many major hospitals, which have been useful for research and quality assurance within these institutions. Although applying microcomputer technology to a statewide system is a promising opportunity, even registries which include data from all the major hospitals in a region will miss cases which are not referred to these centers, and may therefore be inadequate to describe a population which extends over a wide geographic area. Even if the investment of money and time for trauma registry development at smaller hospitals could be justified, this approach would still exclude victims who do not survive to reach a hospital, as well as other cases not requiring admission [1].

The use of existing statewide databases may be a cost-effective method of injury surveillance [2]. However, data collected for other purposes may be less reliable and may not become available for some time after the occurrences of interest. Furthermore, these data may be stored in many different formats, often without direct identifying names or numbers, and it may be difficult to determine which cases are included in more than one source.

We were able to combine registry and population-based data for a study of major burns in Maine, a population small enough to allow matching by human inspection and judgement [3]. It was apparent that a computer would be required to evaluate the much larger population of trauma patients in general. The purpose of the present study was therefore to develop a simple, computer-based method for combining trauma registry data with population-based sources, in order to determine whether this might be a useful method of ongoing data collection in a rural trauma system.

MATERIALS AND METHODS

Programming and data storage have been performed using an IBM Model P75-486 microcomputer (International Business Machines, Armonk NY) using the IBM Personal Computer Disk Operating System (PC-DOS, Version 5.0). Paradox 4.0 (Borland International, Scotts Valley CA) has been used as a database management system (DBMS), and all programs were written in Paradox Application Language or in TurboPascal 6.0 (Borland).

Sources of Data

Hospital trauma registries at the Maine Medical Center (MMC) in Portland and the Eastern Maine Medical Center (EMMC) in Bangor have been maintained using Trauma One (Lancet Technologies, Woburn, MA), a proprietary trauma registry program based upon FoxPro (Fox Software, Perrysburg, OH) which stores data in multiple .DBF files. The use of registry data for

0195-4210/92/$5.00

this study was approved by the Institutional Review Board at each hospital.

A trauma tracking study was developed through the voluntary cooperation of other Maine hospitals, as approved by state legislation. Data have been abstracted from patient records by volunteer nurses, and entered into a single Paradox table designed for this purpose.

Death certificate information for all cases with International Classification of Diseases, Ninth Revision (ICD-9) cause of injury codes corresponding to acute trauma (E800-E848.9, E880-E926.9, E928-E928.9, E955-E979.9, or E985-E999) was requested from the Maine Office of Data, Research, and Vital Statistics, after agreeing to its usual confidentiality restrictions. Computerized abstracts of the death certificate in American Standard Code for Information Interchange (ASCII) file format were provided.

Discharge abstract information for all cases with ICD-9 diagnosis codes corresponding to trauma (800-929.9, 940-959.9) was requested from the Maine Health Care Finance Commission, subject to its usual confidentiality restrictions. Computerized data in ASCII format were provided. These records do not give any direct patient identifying information, and do not distinguish between acute admissions or readmissions of the same patient.

Ambulance run report data were requested from Maine Emergency Medical Services (MEMS) for all cases identified by ambulance personnel as "major trauma", subject to its usual confidentiality restrictions. Computerized data in .DBF format were provided.

Separate programs* were written for each source which extract the data elements of greatest interest (see Table 1) and convert them to a standard ASCII array of characters. This string was then appended to the original file in each case. In the case of ambulance run report data, separate files were created for interhospital transports and for prehospital transports.

Linkage of Data

The main program* reads each of the standardized ASCII data files, matches its records to the other files, and combines the data contained in one or more source for each distinct case. The resulting ASCII file is imported into a Paradox table. This table now contains the most reliable information available for the data points given in Table 1, along with identification numbers or pointers to further data if available from each separate source. A fuller description follows:

The order in which data files are read (Table 2) is based upon the expectation that data collected specifically for injury surveillance will be most reliable, the importance of identifying the most seriously injured cases (deaths and

Table 1: Data abstracted from each source (if available) and placed in the merged table. The first ten fields are also used for matching records from one source to another.

Date of first hospital admission
Sex
Age
First hospital
First hospital length of stay
First hospital disposition
Second hospital, if any
Second hospital length of stay
Second hospital disposition
Date of disposition from last hospital (or date of death if no hospital)
County of injury
Work-related (Y/N)
Best Glasgow Coma Score
Ventilated (Y/N)
E-Code (ICD-9 cause of injury code)
Diagnoses (up to 10, ICD-9)
Identification number (if any)

Table 2: Sources of data, in order read into linkage:

1: Trauma registries at largest hospitals
2: Trauma tracking study
3: Interhospital ambulance run reports
4: Hospital discharge abstracts (for first hospital)
5: Hospital discharge abstracts (for second hospital)
6: Death certificates
7: Prehospital ambulance run reports

Source code available from author.

hospitalizations), and the particular interest in interhospital transfers. The first three sources are able to make positive identification of a transfer between two hospitals. Discharge abstracts may refer to a first hospital or a second hospital. Death certificates will only refer to the patient's final hospitalization, and prehospital ambulance reports must refer to the first hospitalization.

A master file is initialized as empty. For each data source, the first ten fields from each record are read into a temporary data structure consisting of an array of linked lists [4]. Each list represents a "bucket" determined by the patient's age and initial hospital. The master file is then read and each of its records is compared to the new source file in temporary storage; if a match is found, the record in the data structure is marked accordingly. The two files are then read simultaneously from beginning to end and combined into a temporary file, adding the data from marked records in the new source to the appropriate record in the master file, and adding unmarked records as new cases. The temporary file then become the new master file and the process is repeated until all the data files have been read.

This general algorithm is modified for the hospital discharge abstract data, allowing records to be identified as corresponding to a first hospitalization or a second hospitalization. This source must also be compared to itself to identify transfers not contained in previous sources. From this point forward, discharge abstracts identifying first hospitalizations and those identifying second hospitalizations are treated as separate data sources.

The use of "buckets" as described above greatly improves the matching process, requiring only a few short lists to be searched. The identification of possible interhospital transfers requires searching lists corresponding to the other hospitals to which a patient may have been transferred; an empirical function based upon knowledge of local transfer patterns is used for this. The maximum allowable discrepancy in age between two sources can be adjusted, first searching the list of records with no discrepancy, and then lists with progressively larger discrepancies. Allowable discrepancies in dates and lengths of hospitalization can also be varied.

Matched records are required to have the same sex, first hospital, first hospital outcome, second hospital (if any), and second hospital outcome (if any). They must be within the allowable discrepancies for age, date of first hospital admission (except that death certificates are matched by date of hospital disposition), and lengths of stay (if available). Discharge abstracts may also be linked with each other if the first hospital outcome is a transfer, the date of second hospital admission is sufficiently close to the date of first hospital discharge, and the transfer is geographically logical as described above.

Merger of Data

For each data element in Table 1, the sources are ranked in order of their reliability. When matches have been identified and the master table is being updated, each item is examined and changed if the data in the new source has been ranked as more reliable than the existing data. For example, if a patient has data from a hospital discharge abstract and a prehospital ambulance report, the diagnoses will be taken from the discharge abstract, while the location of accident will be taken from the ambulance report.

Due to time and memory limitations, it is not practical to merge and update every possible data field in this way. Furthermore, the discrepancies between two sources may be of interest for certain research questions. Therefore, where direct identifying numbers are available for individual sources, these are also kept in the master file so that they can later be used by end-users to access the individual data bases.

After all iterations of the matching routine have been completed and the master file is in its final form, a new delimited ASCII file is created which expands the coded data into readable text ready for import into a standard database system such as Paradox. As part of this process, the ICD-9 codes are compared with a file provided with the Centers for Disease Control Trauma Registry [5] to obtain Abbreviated Injury Scores; an Injury Severity Score [6] is then calculated. The finished data file is imported into Paradox to facilitate human review.

After computer matching is complete, a series of reports can be generated for human inspection in order to catch errors which could not have been anticipated by the computer. Attention is directed primarily at deaths and interhospital transfers. Errors in the input data can be corrected interactively, and the matching program repeated as many times as necessary to satisfy both human and computer.

RESULTS

Linkage of the various sources of data was thus achieved as planned. The sizes of the data files ranged from 259 records in the Tracking Study to 11645 hospital discharge abstracts. Data availability ranged from immediate, in the case of major hospital trauma registries, to fifteen months after the end of 1991, in the case of death certificates. The cost of obtaining data from state agencies was less than $100.

Computer hardware and software were adequate for a project of this size, although the memory limitations of TurboPascal running under PC-DOS required that the larger files be split in two. The main program as described required about a half hour to process data for one year. Paradox was easily learned by clinicians and staff, and provision of data as general-purpose database tables enabled them to initiate arbitrarily detailed queries.

Establishing that records from different sources truly referred to the same person was not possible, since the use of direct identifying data for most of these sources was not legally permitted and these data were not made available. Nevertheless, using the indirect identifying data described resulted in matching which appeared appropriate by human review of output reports. Once the algorithms had been developed, failures to match were apparently due to coding errors by the Trauma Registrars or state agencies, most commonly involving failure to identify a hospitalization or ambulance run as involving an interhospital transfer.

Allowing the discrepancy in length of stay to exceed one day produced few additional matches regardless of the allowable discrepancies in age and dates. When the allowable length of stay discrepancy was kept constant at one day, increasing the allowable discrepancies in age or dates from one day to two days increased the number of matches by about 6%, but it seemed prudent not to increase the probability of inaccurate matches by doing so.

DISCUSSION

We have chosen hardware and software to be at the same time powerful, flexible, and easily learned, with the intention that the most important end-users will include medical and paramedical professionals who are not trained in computer programming. Paradox was chosen for a database management system because it was highly rated both for quality and for ease of learning [7,8], an assessment with which we strongly agree. For similar reasons, TurboPascal was used for processing text files and PC-DOS for an operating system.

Computer applications specifically designed for trauma data collection and analysis have been developed, and may be helpful within individual hospitals or among hospitals which agree to use the same system. However, in our situation, the variety of formats for existing data required the use of a general-purpose, programmable DBMS. Furthermore, we found that the use of a standard, well-documented DBMS greatly improved our ability to initiate queries and reports. Many good DBMS products are now commercially available, and once data have been put in a standard form they may be easily transferred from one DBMS format to another directly or via ASCII. The skills gained from a modest investment of time learning to use a general-purpose DBMS are also transferrable to other projects.

More sophisticated methods of record linkage have been described [9,10], but even these rely upon data preprocessing, familiarity with local factors, and human review. The use of various statistical or artificial intelligence techniques may be useful for larger databases, but seemed unnecessary for this project. The deterministic program described here is simple enough to be understood after an introductory college computer science course. The chief remaining difficulty is the detection of occasional coding errors, which may require knowledge of the geography, patient variables, or data entry methods to make an appropriate decision about matching.

We do hope to have the opportunity to validate our method against another record linkage program in the near future. If the simpler program is equally effective, it may be less expensive and easier for others to modify and use. Such a comparison may also help establish optimal allowances for date and age discrepancies. The general need for simplifying the record-linkage process has been emphasized [11].

Our greatest challenge will be to go back to the numerous sources which have been linked to determine where and why they are incomplete, with reference to original records where these are available. One outcome of our earlier study was the realization that any single source of data for injured patients misses a significant portion of the total population [3]. The combination of multiple

sources has the potential to provide a more accurate denominator.

Despite the many limitations described, the use of existing sources of data avoids duplication of the effort involved in manual data entry and thereby greatly reduces the cost of data collection. After an initial investment in programming, a computer can rapidly and repeatedly transfer information from one hospital or regional database to another. As the value of this approach is recognized, many of the specific deficiencies of individual data sources can be corrected to make the process more efficient, timely, and complete.

We intend to feed back the data obtained inexpensively through these sources to MEMS and participating hospitals, which can then provide fuller detail in cases of particular interest for system development. Likewise, state agencies responsible for maintenance of other databases can also be alerted to possible procedural errors in order to improve their data. Through an iterative process, a distributed database of continually increasing quality could be possible for injury surveillance and epidemiologic study.

Our goal is a computer database available to clinicians, researchers, and administrators with enough training to initiate basic queries using any standard database management software on a microcomputer [12]. Different levels of access will be required, so that general information can be made available to the public, while patient identifiers used for linkage or quality assurance can be restricted to those having demonstrable need and accountability for this more sensitive information. For research in greater depth, links to more detailed tables can be provided, in a format compatible with any of the more powerful DBMS products. The methods by which records from one source have been matched to another should be accessible for research or system maintenance, but invisible for the great majority of users who will have no interest in this aspect.

Acknowledgements

This work was supported in part by grants from the Maine Health Care Finance Commission and the U.S. Health Resources and Services Administration. The author is indebted to numerous professional colleagues involved in data collection at Maine hospitals and state agencies, especially MMC, EMMC, and MEMS.

References

[1]. A. Cooper, B. Barlow, L. Davidson, J. Relethford, J. O'Meara, L. Mottley. Epidemiology of pediatric trauma: Importance of population-based statistics. J Pediatr Surg 1992; 27:149-154

[2]. T. J. Esposito, J. Nania, R. V. Maier. State trauma system evaluation: A unique and comprehensive approach. Ann Emerg Med 1992; 21: 351-357

[3]. D. E. Clark, M. S. Katz, S. M. Campbell. Decreasing mortality and morbidity following the institution of a statewide burn program. J Burn Care Rehab 1992; 13: 261-270

[4]. E. Horowitz, S. Sahni. Fundamentals of Data Structures in Pascal. New York: Computer Science Press, 1990, pp 79-239

[5]. D. A. Pollock, P. W. McClain. Report from the 1988 Trauma Registry Workshop, including recommendations for hospital-based trauma registries. J Trauma 1989; 29: 827-834

[6]. S. P. Baker, B. O'Neill, W. Haddon Jr., W. B. Long. The Injury Severity Score: A method for describing patients with multiple injuries and evaluating emergency care. J Trauma 1974; 14: 187-196

[7]. J. Duncan, Z. Banapour, N. Petreley, L. Slovick. Product comparison: Data handlers. InfoWorld 12 November 1990; 12: 115-147

[8]. D. Kalman. 15 relational databases: Easy access, programming power. PC Magazine 28 May 1991; 10: 101-200

[9]. L. L. Roos Jr., A. Wajda, J. P. Nicol. The art and science of record linkage: Methods that work with few identifiers. Comput Biol Med 1986; 16: 45-57

[10]. M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Amer Statistical Assoc 1989; 84: 414-420

[11]. A. Wajda, L.L. Roos. Simplifying record linkage: Software and strategy. Comput Biol Med 1987; 17: 239-248

[12]. O'Leary M. Databases of the nineties: The age of access. Database 1990; 13: 15-21